Gamma-Ray Burst Classification: New Insights from Mining Pulse Data

Stanley McAfee and Jon Hakkila

College of Charleston

Despite being the most energetic electromagnetic explosions in the universe, gamma-ray bursts (GRBs) are still poorly understood. The literature recognizes two potentially different types of GRB progenitors, although statistical data suggest the existence of three GRB classes. Reliable inference of GRB physics depends on the identification of appropriate classification attributes, as well as on the statistical classification techniques used. It has recently been shown that pulses are the basic unit of GRB emission. We use new data describing GRB pulse characteristics, in conjunction with data mining tools, to provide a more reliable gamma-ray burst classification system and place additional constraints on GRB physics. We demonstrate that fewer pulses are needed to describe GRB emission than has been suggested by previous analyses, and find pulse duration to be one of the greatest delineators between GRB classes.

I. Introduction

Gamma-ray bursts (GRBs) are brief emissions of high-energy photons lasting tenths to tens of seconds. First detected by satellites in the 1960s, these enormously energetic events have since been confirmed to be isotropically distributed across the sky and cosmological in origin, making them the most powerful electromagnetic explosions in the universe[1]. Broadly, GRBs have two components: an initial flash of gamma-rays called the *prompt emission*, and a lower-energy afterglow that persists following the burst[2]. Here, the *burst* refers to the totality of the original emission event, between where the count rate of detected gamma-rays rises above and returns to the background level.

A representative GRB light curve, detected by the Burst and Transient Source Experiment (BATSE) aboard the Compton Gamma-Ray Observatory, can be seen in Figure 1. The BATSE instrument offers several advantages over modern instruments in the study of GRB prompt emission, including its large surface area and energy range — 20 to 600 keV — allowing it to study GRBs at a variety of signal-to-noise ratios (SNR)[4]. Over the course of its mission, BATSE detected thousands of GRBs. The dataset it produced forms the basis of this work.

While there can be significant variations between GRB light curves, all bursts share a common pulse structure in their prompt emission. The *pulse* is the fundamental component of burst prompt emission, and these pulses share many common characteristics[3,4]. In general, pulses have an *asymmetry,* taking longer to decay than to rise, and evolve from hard to soft energies over time. Softer, lower-energy pulses typically have longer durations and lower asymmetries than higher-energy pulses[5].

Traditionally, a pulse has been defined as a monotonically increasing and decreasing structure that is localized in time. According to this definition, every statistically significant fluctuation that occurs in a GRB light curve can be said to be a distinct pulse[6]. Until recently, this has been the dominant interpretation. However, GRB pulses are *non-monotonic,* exhibiting non-random variations in intensity[5]. This feature of nonmonotonicity is not merely semantic, but rather is essential to understanding GRB pulses, as their properties depend on how they are defined. For instance, if three monotonic pulses are said to describe a GRB light curve instead of one non-monotonic pulse, then the properties of the non-monotonic pulse are effectively distributed over the three monotonic pulses, leading to incorrect conclusions about pulse behavior.

The most obvious example of nonmonotonicity in GRB pulses is a triple-peaked structure that can be seen by fitting a pulse model to GRB data and subtracting out the model to leave the *residuals* of the pulse model. The energy spectra of GRB pulses re-harden at or prior to their peaks[5]. As such behavior would not have been able to be associated with any individual GRB pulse under the old paradigm of monotonicity, this further validates the idea that GRB pulses are non-monotonic.

Pulse and Residual Models

Although it is convenient to describe GRB pulses as having a triple-peaked *residual* structure, as this is where the structure is most readily visible, this is not quite accurate, as it is in fact the *pulse itself* that exhibits the triple-peaked structure. In other words, it is important to distinguish the GRB pulse structure, which is non-monotonic, with the *model* used to describe that structure, which is monotonic. Because it is not known *a priori* what a GRB pulse will look like (and because how the pulse presents itself is a function of factors such as SNR), nonmonotonicity is not included in the pulse model and is instead characterized by the residuals of the model, which together form the "true" shape of a GRB pulse that can be seen in Figure 2[6].

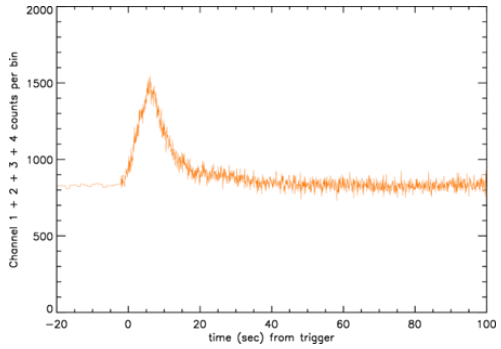

Figure 1 The light curve of BATSE Trigger 06303, a representative single-pulsed GRB.

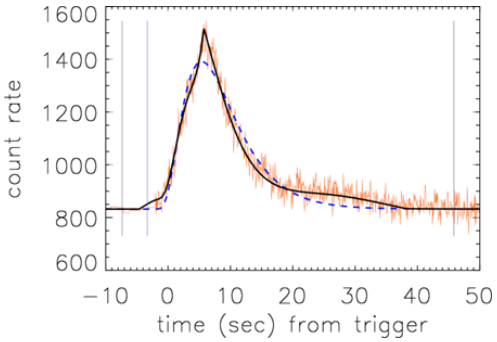

Figure 2. A single GRB pulse. The dotted line is the Norris pulse fit and the solid line is the Norris pulse fit plus the Hakkila -Preece residual fit.

The Norris pulse model pictured in Figure 3 is given by

$$I(t) = A\lambda\, e^{\left(\frac{-\tau_1}{t-t_s} - \frac{t-t_s}{\tau_2}\right)}$$

where A is the pulse amplitude, t is the time elapsed since the trigger event, t_s is the pulse start time, τ_1 and τ_2 are respectively the pulse rise and decay parameters, and $\lambda = e^{2\sqrt{\tau_1/\tau_2}}$ is the normalization constant[14].

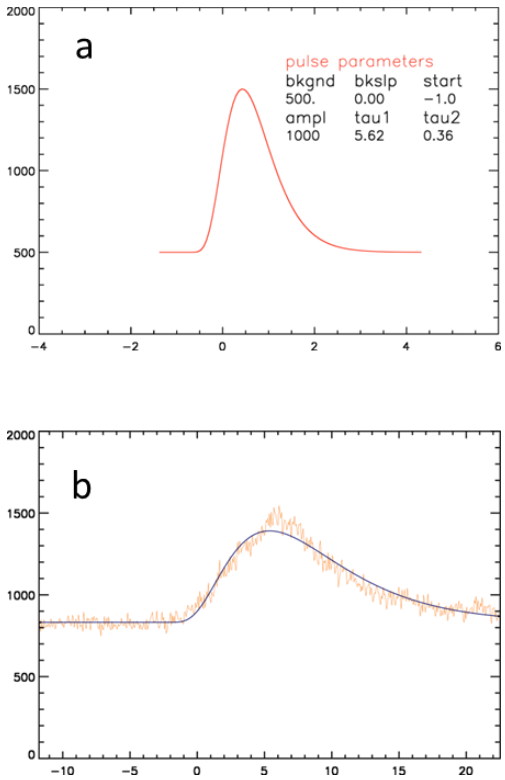

Figure 3. The Norris pulse model. The x axis is the time since the trigger event and the y axis is the number of counts in each bin. Figure 3a. The monotonic Norris pulse model. Figure 3b. The pulse fit to BATSE Trigger 06303.

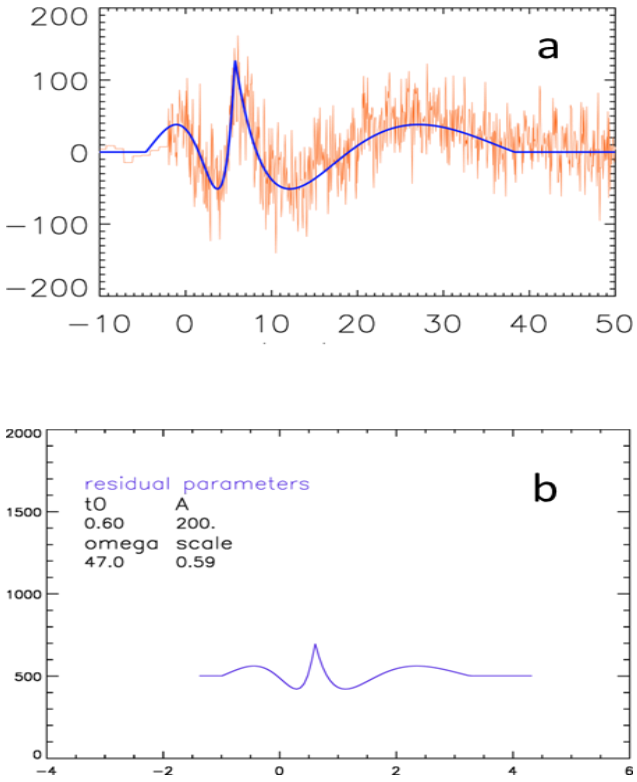

Figure 4. The Hakkila-Preece residual model. The x axis is the time since the trigger event and the y axis is the counts in e ach time bin. Figure 4a. The Hakkila-Preece residual model. Figure 4b. The residual fit to BATSE Trigger 06303.

The Hakkila-Preece residual model seen in Figure 4 is given by

$$res(t) = \begin{cases} AJ_0\sqrt{\Omega(t_0 - t - \Delta/2)}, & t < t_0 - \Delta/2 \\ A, & t_0 - \Delta/2 \le t \le t_0 + \Delta/2 \\ AJ_0\sqrt{s\Omega(t_0 - t - \Delta/2)}, & t > t_0 + \Delta/2 \end{cases}$$

where J_0 is an integer Bessel function of the first kind, t_0 is the central time of the peak amplitude, A is the normalized amplitude of the peak, Δ is the duration of the peak, Ω is the Bessel function's angular frequency, and s is a scaling factor[5].

II. Motivation

While the mechanisms that produce GRBs remain a mystery, there exist commonalities between different bursts. Three GRB classes have been found on the basis of burst properties such as duration, fluence, and spectral hardness: Short, Long, and Intermediate bursts. Short bursts have shorter durations, lower fluences, and softer spectra than Long bursts; Intermediate bursts have intermediate durations and fluences as well as soft spectra[7,8]. The delineation between a Short and a Long burst has traditionally been made on T_{90} duration, the time it takes to accumulate 90% of the fluence of a burst, based on a bimodality in the logarithmic durations of GRBs observed by the BATSE instrument. As its range of durations straddle this divide, shown in Figure 5, the existence of an Intermediate class has primarily been indicated by statistical and data mining techniques. While different methods recover different properties – and, in some cases, different numbers of classes – the recovery of a third class at all suggests that the current burst classification scheme is incomplete.

Classification schemes have also been developed for GRB pulses on the basis of correlated pulse properties such as spectral lag, duration, and

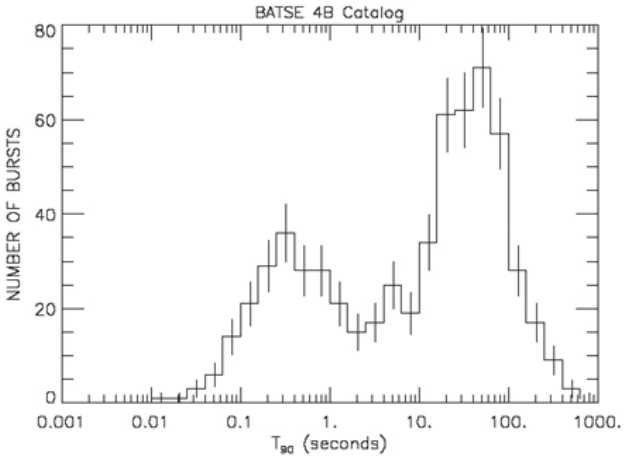

Figure 5 The GRB logarithmic duration bimodality observed by the BATSE instrument, occurring at a $T90$ of $\approx 2s$. Image from icecube.wisc.edu/~ms25/T90_distribution.jpg.

asymmetry. Two classes of GRB pulses have been identified through this process: short, spectrally hard, symmetric, short-lag pulses, and long, spectrally soft, asymmetric, long-lag pulses. These pulse classes are associated with the Short and Long GRB classes respectively, with pulses in multi-pulsed GRBs having less distinctive properties than those in single-pulsed GRBs[9]. Within the Short burst class, pulses have also been classified based on their *complexity*, defined in relation to differences in the X^2 values between the Norris pulse fit and the Norris pulse fit plus the Hakkila-Preece residual fit. As compared to Long and Intermediate GRBs, pulses in Short GRBs were found to be spectrally harder and more likely to be single-pulsed[6]. While clear delineations were discovered between pulses in Short and Long bursts in all cases, the same was not true for Intermediate pulses, which were largely

separated by spectral hardness.

As past efforts to classify GRBs have primarily been done on the basis of *burst* properties, the central question of this research is whether or not these classes can be recovered from *pulse* properties as well. Classifying GRBs can provide insights into, and constraints on, the physical mechanisms that produce them, and characterizing the properties of GRB pulses may be able to do the same for pulse models.

III. Analysis

A. Pulse sample

A sample of pulses was produced for analysis with data mining tools by fitting pulses to sequential bursts in the 1000s trigger group of the BATSE dataset. Bursts were excluded if their data was incomplete or otherwise contaminated (e.g. a particle event or solar flare), and care was taken to ensure the sample was as complete and unbiased as possible. The details of how this was accomplished are discussed in Section IV.a.

Table 1 lists the BATSE trigger IDs of the bursts analyzed, as well as their published burst class and the number of pulses fit to each burst. 28 bursts were fit, resulting in a sample of 31 pulses.

1. Pulse Identification and Fitting

The Bayesian Blocks algorithm was implemented in Interactive Data Language (IDL) programs to identify pulses in GRB light curves. Conceptually, the algorithm functioned by dividing the data into regions (the eponymous Bayesian Blocks), searching within those regions for statistically significant variations according to user-specified criteria for statistical significance, and iterating through this process until all insignificant variations were culled. Figure 6 shows the outcome of this process when applied to a representative GRB light curve.

Table 1 The bursts analyzed in this work, along with their associated classes and the number of pulses fit to each burst. Burst classes from Hakkila et. al. 2018[6].

Burst ID	Class	Num pulses
1039	Long	1
1042	Long	3
1051	Short	1
1073	Short	1
1076	Short	1
1085	Long	1
1086	Long	1
1088	Short	1
1096	Short	1
1097	Short	1
1102	Short	1
1110	Long	1
1112	Short	1
1114	Intermediate	1
1120	Long	1
1125	Long	1
1126	Long	1
1129	Short	1
1141	Long	1
1145	Long	1
1148	Long	1
1153	Long	1
1159	Long	1
1167	Long	1
1190	Long	1
1196	Long	2
1200	Long	1
1211	Short	1

Once a potential pulse had been identified by the Bayesian Blocks algorithm, the Norris pulse model was fit to the data using an IDL implementation of **MPFIT,** a popular least-squares curve fitting routine. **MPFIT** iteratively varied the parameters of the Norris model until a minimum value of X^2 was reached, which served as a test of goodness-of-fit. The **MPFIT** routine was also used to fit the Hakkila-Preece residual model to the residuals of the pulse fit using the same procedure.

How many pulses to fit to the data is a very important question, particularly given the previous discussions of GRB pulse nonmonotonicity and sample completeness and bias. Even if a GRB pulse is understood to be non-monotonic, it is rarely clear beforehand to what extent the nonmonotonicity of GRB light curves is associated with the underlying pulse, and as we only see GRBs after they cross cosmological distances and interact with a detector, deconvolving the pulse from these other (potentially unknown) effects is challenging and becomes even more so when the requirement of statistical significance is added. It is simple enough to make qualitative observations about GRB behavior, but difficult to justify these observations in a quantitative way.

The Bayesian Blocks algorithm provides this sort of statistical justification for identifying a pulse, but it can identify an arbitrarily large number of pulses per burst if improper statistical criteria are specified. Even worse, the X^2 statistic can appear to improve with increasing pulse number, because the fit can become almost exact. In this way, the principle of Occam's razor (coupled with the assumption of pulse nonmonotonicity) is a useful constraint on the Bayesian Blocks algorithm. If every variation in a GRB light curve is not considered to be a pulse, then why fit ten pulses to a burst instead of five? Why fit five instead of three, or even one? "The minimum number of pulses necessary to account for the behavior of the burst" is a reasonable answer to the question of how many pulses to fit, but it raises the related question of what exactly that minimum number is. In other words: how do you know you've fit the right number of pulses to the data?

2. Temporal Rebinning

One method used in this work to determine the number of pulses to fit to an individual GRB, particularly those whose light curves exhibit complex variations, was temporal rebinning of the residuals of the pulse fit to larger timescales. Previous work by Eric Hofesmann has indicated that the summed residuals of complex GRBs contain the triple-peaked residual structure characteristic of a single pulse[13]. This suggests that, however complicated and variable their structure, an individual complex GRB may itself contain the triple-peaked structure characteristic of a single pulse--- and that if this is the case, it is justified to fit the emission episode with a single pulse and compare the results of this process to other non-complex, single-pulsed GRBs. If the pulse is in some sense "hiding" beneath the complex variations, temporally rebinning the residuals to larger and larger timescales can illuminate the underlying triple-peaked structure and thus the underlying pulse.

This technique was applied to an artificial sample GRB pulse in order to determine an upper limit on the binsize before the residual

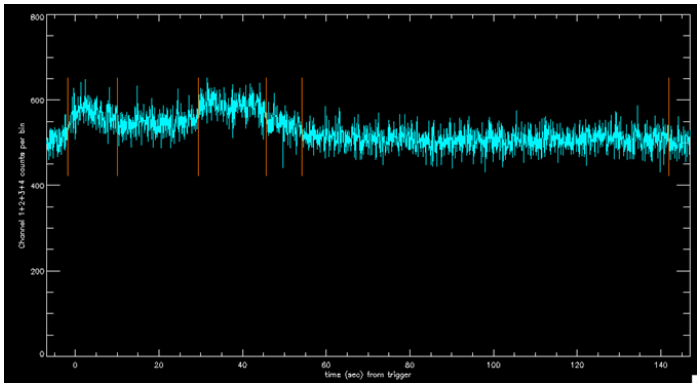

Figure 6. Bayesian Blocks displayed on a GRB light curve. The jagged blue line is the data and the vertical orange lines indicate the locations of Bayesian Blocks.

structure is no longer recoverable. The effects of asymmetry, residual to pulse amplitude ratio, and bin shifting before rebinning were taken into account, and a $|\Delta X^2|$ value was produced for each increasing binsize to determine when the technique no longer yielded useful results. Though the results of this study are not directly applicable to complex GRBs, as the sample pulse and its residuals were well-behaved, it provides a rough upper limit on how large the time bin can be as a fraction of total pulse duration before only noise is left behind. Examples of these results are shown in Figure 7.

Two different complex bursts, GRBs 143 and 1114, provide evidence for the efficacy of temporal rebinning to determine the existence of an underlying single pulse. Of these, GRB 143, pictured in Figure 8, offers the clearest justification for the application of this technique. While the pulse fitting code reliably fits more than one pulse to the first emission episode, the emission episode itself has the proper asymmetry and "emission envelope" of a single pulse. Furthermore, the second emission is much better-behaved, showing the clear triple-peaked structure of a single pulse. As the second emission is related to the first, it is possible that the first is in fact a single pulse as well. The results of fitting a single pulse to both emission episodes and subsequently rebinning their residuals to larger timescales are seen in Figure 9.

The light curve of GRB 1114, shown along with the pulse fit in Figure 10, demonstrates fewer behaviors associated with a single pulse. However, given the previous success of the technique when applied to GRB 143, it is reasonable to treat the emission episode as a single pulse and rebin the residuals of the fit in order to see if they contain a triple-peaked structure. This process is illustrated in Figure 11. By Figure 11 c, the triple-peaked structure is recovered, suggesting that a single GRB pulse underlies the complex behavior of the light curve.

B. Attribute Selection

Once the pulse sample was produced and the properties of each pulse extracted to a database, a subset of these properties were selected as

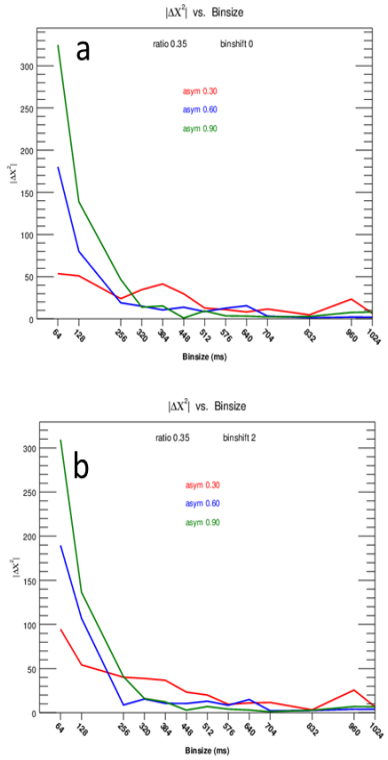

Figure 7. $|\Delta X^2|$ values vs. binsize for a 5s duration sample pulse. The red, blue, and green lines correspond to low, medium, and high asymmetries respectively. Larger $|\Delta X^2|$values indicate greater improvement after including the rebinned residual fit to the pulse fit. Figure 7a. No bin shifting before rebinning. Figure 7b. Shifting two bins before rebinning.

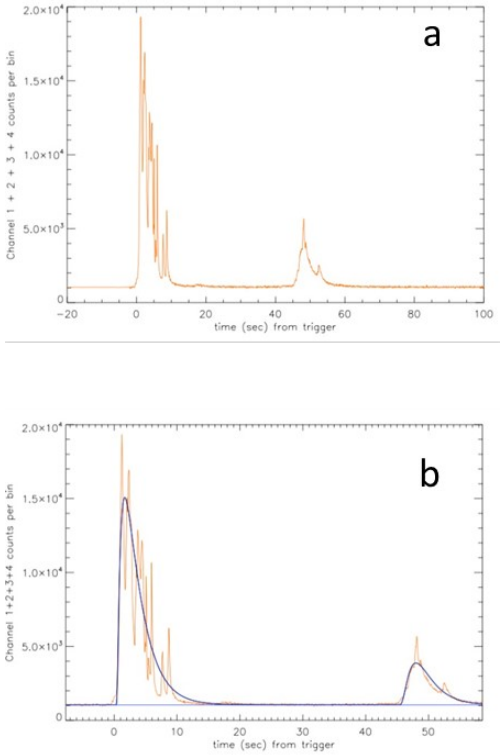

Figure 8. Light curves and pulse fits of GRB 143. Figure 8a. The light curve of GRB 143. Figure 8b. A single pulse fit to each emission.

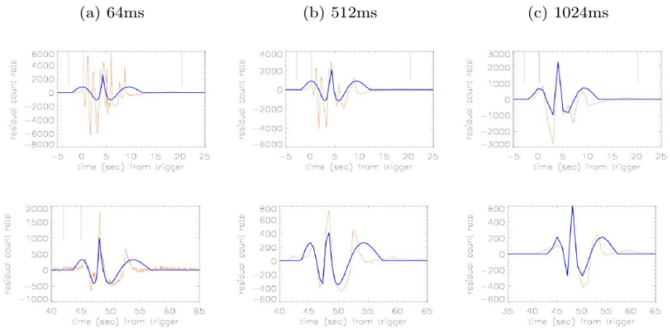

Figure 9 The effects of rebinning the residuals of GRB 143 to larger timescales. The top row is the first emission and the bottom row is the second emission. At a binsize of 1024ms, the complex first emission shows evidence of the triple-peak residual structure.

classification attributes. Useful classification attributes are those that avoid bias, irrelevance, redundancy, and over-specification while also sufficiently characterizing the sample. The objective of attribute selection is to determine which attributes contribute useful information to the classifiers and which do not.

The Waikato Environment for Knowledge Analysis (Weka) data mining suite contains a variety of attribute selection tools called *evaluators*. Each evaluator uses different criteria to determine usefulness. Three different evaluators were used: **CfsSubsetEval**, which considers the individual predictive ability of each attribute as well as the degree of redundancy between them; **InfoGainAttributeEval**, which evaluates the worth of an attribute by measuring the information gain with respect to the class; and **CorrelationAttributeEval,** which evaluates the worth of an attribute by measuring the Pearson's correlation between it and the class. The attributes that were determined to be the most useful by all three evaluators were selected as the final classification attributes, and are recorded in Table 3 along with their

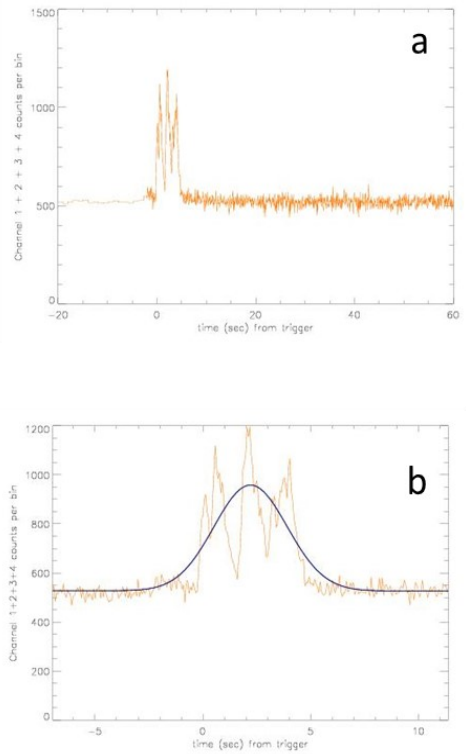

Figure 10. Light curves and pulse fits of GRB 1114. Figure 10a. The light curve of GRB 1114. Figure 10b. A single pulse fit to the emission.

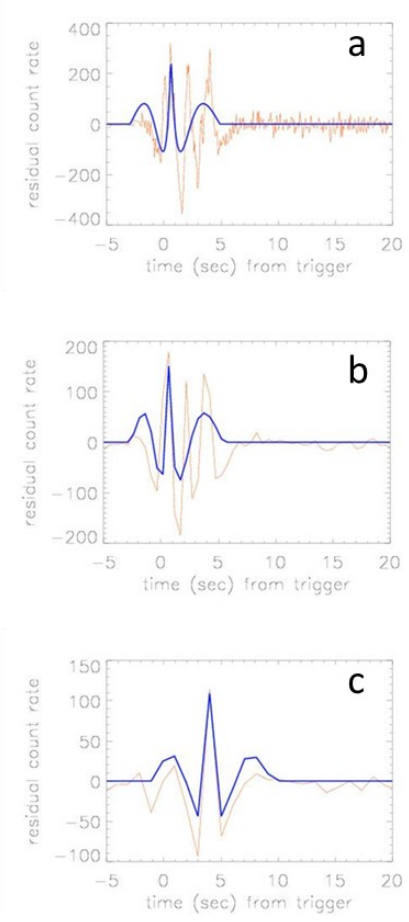

Figure 11. The residuals of GRB 1114, rebinned to larger timescales. Figure 11a 64ms residuals. Note the position of the residual peak here as compared to its position in the 1024ms residuals. Figure 11b 512ms. Figure 11 c 1024ms

hypothesized relationship to class. The attributes that were not useful are recorded in Table 2 along with the reason for their removal.

C. Clustering

The process of *clustering* seeks to identify similarities between different objects and group (or *cluster*) them according to those similarities. These clusters are most readily visualized in "attribute space," where every attribute of the dataset represents a different axis. Related objects appear to be clustered near each other when pictured in this way. The concept of attribute space also helps to explain the importance of attribute selection. If each axis is not strictly orthogonal — that is, if there is redundancy between attributes — or if some axes are irrelevant, it is more difficult to identify unique clusters of data.

As with attribute selection, Weka contains a number of clustering algorithms to perform clustering analysis. Different clustering algorithms have different approaches to the problem of clustering, and make different assumptions about how the data is distributed in order to identify clusters. For most real applications, the true distribution of the data is unknown, as is the nature of any similarities between the data. Even if these were known, there is also no "correct" way to cluster the data, as the value of the clustering scheme depends on the usefulness of the resulting information to the user. In other words, even the crudest attempt at clustering could potentially provide insight into some feature of the data, even if other features are obfuscated.

Therefore, to obtain the broadest picture of the data possible and increase the likelihood of recovering useful information, two clustering algorithms — *K*-Means and *Expectation Maximization* (EM) — were used. Because the distribution of the data can affect the outcome of the clustering process, two different versions of the sample were created: one containing the original attributes (the original sample), and one containing logarithms of the original attributes (the log sample).

1. *K*-Means

The *K*-Means algorithm separates data into *K* clusters, with *K* specified by the user, and assigns data to the cluster with the nearest mean. The result is data clustered about a centroid that defines the cluster. The *K*-Means algorithm has two major parameters: the initial *K* specification, and the method by which the distance between each data point and the centroid is computed. The most intuitive method is to use the Euclidean distance in attribute space. For example, given attributes x, y, and z, the distance between a data point d and the centroid c is

$$\sqrt{(x_d - x_c)^2 + (y_d - y_c)^2 + (z_d - z_c)^2}$$

However, there exist alternative ways of calculating this distance, in particular the Manhattan distance, which uses absolute differences between the individual parameters instead of the overall distance in attribute space. A simple diagram of these distance measures can be seen in Figure 12.

The *K*-Means algorithm was applied to the original sample and the log sample for $K = 2$ and $K = 3$, with both distance measures used for each *K* value. The effectiveness of each combination is recorded in Table 4.

2. Expectation Maximization

The EM algorithm assigns a value to each data point, representing the probability of it belonging to a cluster. These clusters are assumed to be Gaussian distributions. Starting with an initial guess, the algorithm iteratively improves its estimates of the properties of the distributions until the probability that the data point belongs to that distribution is maximized. Unlike *K*-Means, the EM algorithm does not require the number of clusters to be pre-specified, but instead can automatically determine the number of clusters necessary to describe the data. This makes the EM algorithm a useful way to check the results of the *K*-Means algorithm and provide additional justification for the number of clusters selected.

The EM algorithm was applied to the original sample and the log sample, with the results recorded in Table 5.

Table 2 Attributes that were not used for classification, as well as the reason for their removal.

Attribute	Description	Reason for removal
bkgnd/bkslp	Burst background count rate, determined by Norris pulse model fit	Background levels, by definition, have nothing to do with the pulse, making them poor classification parameters
ampl/a/R	Pulse amplitude, residual amplitude, and pulse/residual amplitude ratio respectively	Pulse amplitude correlates with SNR and is not a very reliable parameter at low SNR; if this is true, then residual amplitude is also a suspect parameter, and removing both **ampl** and **a** necessitates the removal of **R**
tau1/tau2/ kappa/taupk	Pulse rise parameter, pulse decay parameter, pulse asymmetry, and time of pulse asymmetry respectively	Individual **tau1/tau2** values for each pulse are not very illuminating, as asymmetry information is contained within kappa, but kappa itself is highly prone to error and correlated with hardness; removing these parameters necessitates removal of **taupk**, which depends on them
dualflx/ fluen_ch1/2/3/4	The dual flux time scale on which the pulse triggered, and the fluences in each energy channel of the detector	The dual flux timescale on which the pulse triggered is interesting information, as it provides insight into the biases of the detector, but not necessarily into the pulses; fluences per channel are useful to determine what percentage of the emission was hard vs. soft, but this information is contained within the hardness ratio, making them duplicitous
All lags except **lag_1_3**	Pulse spectral lags between the high energy and low energy channels, indicating hard to soft evolution	Any lag associated with channel 4 is likely to be biased as a result of fewer fits converging in channel 4; **lag_1_3** was reliably selected as a useful parameter and provides more insight into hard to soft evolution than **lag_2_3**, which did not provide new information over **lag_1_3** according to several attribute evaluation methods
t0	Residual start time	It has been shown that the pulse and residual start times do not necessarily coincide; however, as the "true" (i.e. not modelled) residual structure of the pulse has recently been called into question, it is unlikely that the residual start time is either correct or a useful delineator between classes of pulses
omega/s	Residual angular frequency and scaling factor	Omega and s are correlated, because for a different frequency a different "stretching" value is required to achieve a desired result; this eliminates s, leaving omega, which is already strongly correlated with pulse duration

Table 3 Final classification attributes and their descriptions, as well as their hypothesized relationship to class.

Attribute	Description	Hypothesized relationship to class
num pulses	Number of pulses fit to the burst	Shorter bursts will have fewer pulses than longer bursts
start	Start time to the pulse relative to the trigger time of the burst	Shorter pulses will start more quickly than longer pulses
dur	Duration of the pulse	Shorter pulses will be found in shorter bursts
fluen	Fluence of the pulse	Shorter pulses will have a lower fluence than longer pulses
lag_1_3	**taupk_ch1 – taupk_ch3**, how much longer it takes for the emission to peak in the lower energy channel vs. the higher energy channel	Shorter pulses will have shorter lags than longer pulses
SNR	Signal-to-noise ratio of the pulse	Shorter pulses will have lower SNRs than longer pulses
hardness	Hardness ratio of the pulse, **ch3+ch4/ch1+ch2** counts hardness	Shorter pulses will have higher hardness ratios than longer pulses
1_3_spread	**dur_ch1/dur_ch3**, "spreading" of low energy vs. high energy (low values mean longer duration high energy emission than low energy emission)	Shorter pulses will have smaller spreads than longer pulses

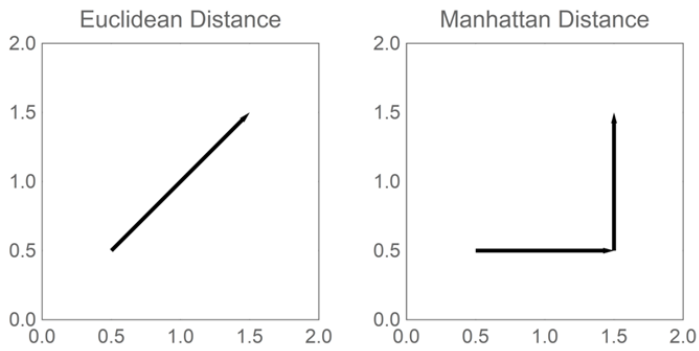

Figure 12 Euclidean distance vs. Manhattan distance measures.

Table 4. The results of applying the *K*-Means algorithm with different options to each version of the pulse sample. In this context, "incorrectly clustered" means a failure to recover the classes in Table 1.

Dataset	*K* value	Distance measure	Incorrectly clustered
Original	2	Euclidean	23%
		Manhattan	13%
	3	Euclidean	23%
		Manhattan	16%
Log	2	Euclidean	13%
		Manhattan	7%
	3	Euclidean	16%
		Manhattan	10%

Table 5.. The results of applying the EM algorithm with different options to each version of the pulse sample. In this context, "incorrectly clustered" means a failure to recover the classes in Table 1. The log-likelihood value is a measure of how likely it is the clusters characterize the data.

Dataset	Incorrectly clustered	Log-likelihood
Original	29%	-22.4
Log	13%	-1.89

3. Comparison and Results

For the original sample, the performance of the algorithms, from most effective to least effective, was

1. 13%, *K*-Means (Manhattan, K=2)
2. 16%, *K*-Means (Manhattan, K=3)
3. 23%, *K*-Means (Euclidean, K=2)
4. 23%, *K*-Means (Euclidean, K=3)
5. 29%, EM

For the log sample, the performance of the algorithms, from most effective to least effective, was

1. 7%, *K*-Means (Manhattan, K=2)
2. 10%, *K*-Means (Manhattan, K=3)
3. 13%, EM
4. 13%, *K*-Means (Euclidean, K=2)
5. 16%, *K*-Means (Euclidean, K=3)

where in both cases the percentage indicates what fraction of the data was "incorrectly clustered" according to the classes specified in Table 1.

Figure 13 shows the results of EM clustering on the log sample data, with every classification attribute plotted against every other. 18 pulses were identified as Long, 10 as Short, and 3 as Intermediate, as compared to the data from Table 1, which contained 20 Longs, 10 Shorts, and 1 Intermediate.

D. Classification

While clustering analysis in some sense produces a classification scheme, these "classes" can be difficult to associate with the attributes used to produce them. It is often unclear what attributes contributed the most to defining the clusters or what physical significance the mathematical definition of the cluster has. A proper classification scheme does more than simply associate similar objects— it explains those associations in the context of class behaviors. Therefore, when performing data mining, *classifiers* are used to explain the results of clustering analysis.

A popular type of classifier is called a *decision tree*. A decision tree uses the values of an object's attributes (the branches) to determine the class it belongs to (the leaves). Depending on the algorithm, different criteria are used to decide when to split a branch. For instance, the J48 decision tree algorithm, implemented in Weka and used in this project, uses a measure of information gain to decide when to split a branch. Decision trees are powerfully simple, producing easy to interpret rules for class membership.

The standard procedure for classifying data is to separate it into a smaller *training* and larger *testing* set. The classifier is first "trained" on the smaller training set, using the minimum amount of information necessary to determine classification rules. Once these rules are created, they are tested on the larger testing set to determine their efficacy. Following this procedure is important: if no distinction is made between the training and testing sets, the resulting classification scheme will be perfectly tailored to the data used to create it, an outcome called overfitting. As a classification scheme is meant to be applied to new observations, this is manifestly undesirable.

A typical training set consists of about a third of the total sample. If the total sample size is small, selecting an even smaller subset can make it difficult for the classifier to produce sensible rules. One way to obviate this issue is to use *n-fold cross validation*, which splits the data into *n* subsets and gives each subset a chance to be part of the testing and

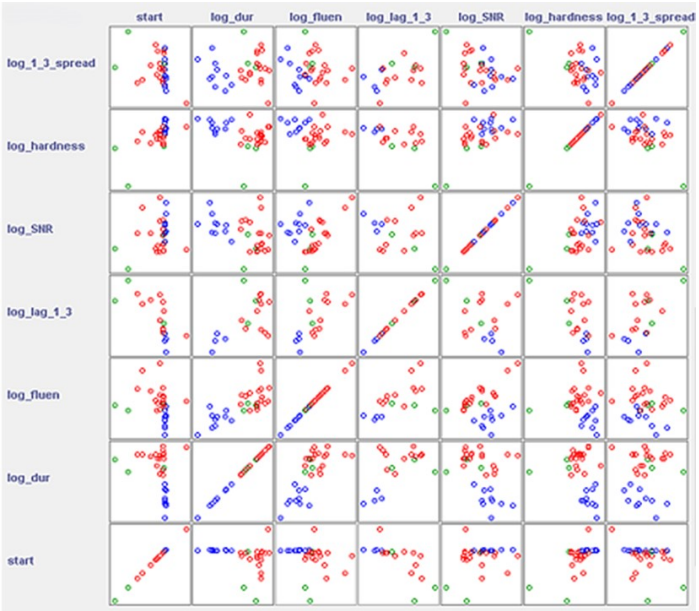

Figure 13 The plot matrix produced by EM clustering of the log sample. Blue indicates the Short class, red the Long class, and green the Intermediate class.

training sets. This generally increases the robustness of the rules generated by the classifier.

J48 classification was performed on the top three most effective clustering results from the log sample: EM, *K*-Means (Manhattan, K=2), and *K*-Means (Manhattan, K=3). Due to the small sample size, 10-fold cross-validation was used, and pruning — the removal of branches of the decision tree that contribute no new information — was disabled so as to produce the largest number of classification rules possible.

The decision tree for the EM clustered data was

```
log_dur <= 0.131384: Short* (10.0)
log_dur > 0.131384
|   log_hardness <= -0.692699: Intermediate*
(3.0)
|   log_hardness > -0.692699: Long* (18.0)
Relative absolute error: 6%
```

where the number in parentheses indicates the number of pulses assigned to this class by the decision tree.

The decision tree for the *K*-Means (Manhattan, K=2) clustered data was

```
log_dur <= 0.580362
|   log_hardness <= -0.426909: Long* (1.0)
|   log_hardness > -0.426909: Short* (11.0)
log_dur > 0.580362: Long* (19.0)
Relative absolute error: 14%
```

The decision tree for the *K*-Means (Manhattan, K=3) clustered data was

```
log_dur <= 0.131384: Short* (10.0)
log_dur > 0.131384
|   log_hardness <= -0.772749: Intermediate*
(2.0)
|   log_hardness > -0.772749: Long* (19.0)
Relative absolute error: 13%
```

In all cases, Short pulses were distinguished from Long pulses on the basis of pulse duration, and Intermediate pulses from Long pulses on the basis of spectral hardness. This distinction can clearly be seen in Figure 14.

Because these rules define class membership, decision tree algorithms require a class to evaluate them against. Here, the classes provided were the clusters produced by clustering analysis, which already did not exactly align with the classes in Table 1. Therefore, the classification errors produced by the J48 algorithm above do not represent misidentification of the classes in Table 1, but instead serve as a coarse measure of how "confident" the classifier was in producing the rules.

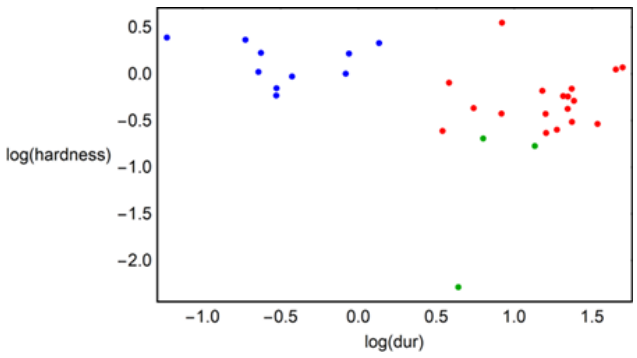

Figure 14 Logarithmic duration vs. logarithmic hardness for EM clustered data. Blue indicates Short pulses, red indicates Long pulses, and green represents Intermediate pulses. All subsequent figures adhere to this convention.

IV. Discussion

A. Sample Completeness and Bias

Data mining tools are sensitive to sample completeness, particular in the context of GRB classification[10]. As such, care was taken to ensure the pulse sample was as *complete* and *unbiased* as possible. *Completeness* refers to how representative the sample is of the full population. As the sample here was created by fitting a model to data, in this context, a complete sample is one that contains examples of both known, regular behavior that is easily characterized by the model as well as unknown, irregular behavior that is not as easily characterized by the model. An incomplete sample, then, does not contain examples of all possible behaviors of the population. *Bias* indicates a systematic difference between the behavior of the sample as the behavior of the population. Therefore, an incomplete sample is always biased, while a complete sample could be biased or unbiased, depending on the validity of the techniques used to produce it.

As outlined in Sections III.a.1 and III.a.2, multiple techniques were used to avoid biasing the pulse fitting process. While largely successful, these techniques were not able to resolve every ambiguous case, particularly for multi-pulsed bursts. Some ambiguity is to be expected when considering these types of events, if only due to the potential for overlapping pulses, but the multi-pulsed bursts present in the 1000s trigger group of the BATSE catalog were exceptionally confusing, as many did not adhere to standard definitions of pulse behavior even with the assumption of pulse nonmonotonicity.

However unusual, these strange types of bursts could not simply be ignored. But this created a dilemma: in order to include in them in the sample, they must first have been fit with the Norris pulse model, but their strangeness almost by definition precluded this. Even so, the pulse fitting process was attempted for every burst, despite being fraught with difficulty. The large number of emission episodes contained within each burst complicated the process of determining the proper number of pulses to fit, and even if the number of pulses to fit to each burst were known, the irregular nature of these pulses – namely, a "backwards" asymmetry that took longer to rise than decay – challenged the assumptions made by the pulse fitting code and resulted in incorrect values and enormous errors for parameters such as the start time and asymmetry. "Forcing" a fit by explicitly defining the locations of the Bayesian Blocks and/or the pulse model parameters was technically possible, but merely substituted one kind of bias with another that was even less statistically justified, and was therefore not considered to be a viable method.

Ultimately, when faced with situations where the Norris pulse model failed to characterize the behavior of the burst, and continued to fail even after manipulating the parameters of the pulse fitting code, the offending burst was excluded from the sample. This had the effect of biasing the sample towards bursts with single emission episodes, which were able to be satisfactorily characterized by the pulse model even if their light curves varied in complex ways, and therefore is incomplete with regards to multi-pulsed events. Even so, that complex GRBs like those in Figure 8 b and Figure 10 b can be fit with a single pulse, and that useful information results from this procedure, indicates a development in the understanding of GRB pulses, at least with respect to how many pulses make up an average burst.

B. Clustering Effectiveness

Logarithmic pulse properties were much more effective for clustering analysis than the pulse properties themselves, particularly for the EM algorithm, which saw a dramatic improvement from -22.4 to -1.89 log likelihood that the data was generated by the parameters provided— that is, that these parameters define GRB classes. Because the EM algorithm assumes a Gaussian distribution, a log-normal distribution of pulse properties would explain its vastly improved effectiveness.

For the *K*-Means algorithm, the Manhattan distance function was superior to the Euclidean distance function in every case. One possible explanation is that the Manhattan distance function tracks the absolute differences between individual parameters, while the Euclidean distance tracks the geometric distance between all parameters in a higher-dimensional space. This makes the Manhattan distance function less

sensitive to outliers in the data and may allow it to identify clusters of individual parameters more readily than the Euclidean distance, which examines the aggregate relationship of all the parameters.

While two clusters was the preferred solution of the *K*-Means algorithm in all cases, the unsupervised EM algorithm routinely identified three clusters, although these three clusters did not necessarily follow the Short/Intermediate/Long classification scheme, particularly with respect to the Intermediate class. The three pulses identified as Intermediate by the EM algorithm were all from BATSE trigger 1042, which was classified as Long in Table 1 and is the only instance of a burst with three pulses in the sample. As discussed in Section IV.a this makes the Intermediate class suspect, as it could have arisen from sample incompleteness. Even if multi-pulsed bursts truly do constitute a separate class on the basis of number of pulses alone, Figure 13 demonstrates that the other properties of the pulses in 1042 – such as start time, lag, and hardness – are at best outliers and at worst entirely inaccurate. Given these results, it is likely that the Intermediate class identified here does not represent a distinct population of pulses.

C. Classification Effectiveness

Due to the low number of multi-pulsed bursts present in the sample, number of pulses was a significant class property for many of the algorithms. However, because this number was so low and because there is a known bias towards bursts containing a single emission event, this an unreliable attribute for classification. Classifying without its inclusion led to more meaningful classification rules on the basis of duration and hardness, relationships identified in previous, larger analyses of burst properties[9 3].

For the data clustered with *K*-Means, the J48 relative absolute error (again, interpreted here as confidence in the classification rules) for $K = 2$ and $K=3$ and were comparable, with $K = 3$ only marginally the preferred solution. The classification rules generated for $K = 2$, however, were less useful than those generated for $K=3$. The $K=3$ classification rules delineated classes on the basis of duration (with a break at $\approx 1.3s$) and spectral hardness, with longer, softer pulses classified as Intermediate and longer, harder pulses classified as Long. These rules are essentially the same as those generated for the data clustered with EM, which had a lower absolute relative error. Therefore, duration and hardness appear to be two of the most important parameters in assigning a pulse to a class.

D. Short and Long Pulse Behaviors

Although the Intermediate class was indistinct, the clear distinction between Short and Long pulses on the basis of duration and spectral hardness indicates that they likely belong to different populations. One method of testing this hypothesis is by examining the plots of SNR in Figure 13. Although a variety of factors contribute to SNR, if classes appear distinct at high SNR, it is reasonable to suppose that those distinctions are real. For example, with increasing SNR, Short pulses become shorter. Similar correlations were observed between SNR and hardness as well as **1_3_spread**, which can be seen in Figure 15.

A correlation ($p = 0.044$) was observed between SNR and hardness for both the Shorts and Longs in Figure 15 a. At larger SNR, pulses become spectrally harder. SNR increases for GRBs from which we receive the most signal, which could occur because the emission is directed along our line of sight or because the event itself is intrinsically brighter. In either case, a greater fraction of higher energy photons should be detected, which is consistent with this observation. Although not statistically significant, a weak anti-correlation was also observed between SNR and **1_3_spread** for both the Shorts and the Longs in Figure 15 b. This further supports the previous argument, as **1_3_spread** should decrease with increasing hardness and thus with increasing SNR.

Within the Short pulse class, a weak anti-correlation ($p = 0.093$) was observed between duration and **1_3_spread** As pulse duration increased, the duration in channel 3 increased relative to channel 1. This can be seen in Figure 16. This relationship is curious, because if overall pulse hardness increases with increasing SNR, **1_3_spread** decreases with increasing SNR, and Short pulse durations decrease with increasing SNR, it seems intuitive to suspect that **1_3_spread** would be positively correlated with Short pulse duration. That is, longer Short pulses ought

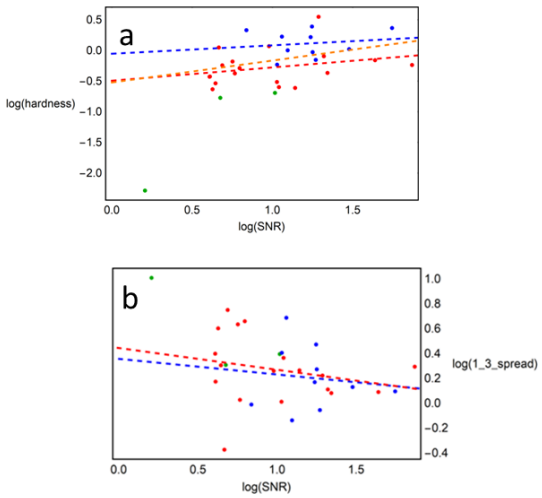

Figure 15. Logarithmic SNR vs. logarithmic hardness and channel 1/channel 3 spread. The red lines indicate a fit to the Long pulses, the blue lines indicate a fit to the Short pulses, and the orange line represents a fit to the Short and Long pulses. Figure 15a. Logarithmic SNR vs. logarithmic hardness. Figure 15b. Logarithmic SNR vs. logarithmic channel 1/channel 3 spread.

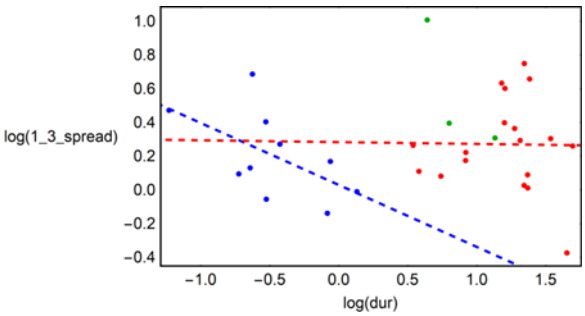

Figure 16 Logarithmic duration vs. channel 1/channel 3 spread.

to be softer overall and therefore have higher values of **1_3_spread**, but this is not the case.

As this relationship was not observed within the Long class, this may indicate a difference in the mechanism producing Shorts and Longs, although it is difficult to draw a firm conclusion due to the low statistical significance of the result. A different, less dramatic explanation is that the durations of the high energy vs. low components of the pulse do not necessarily reflect the overall hardness of the pulse, an observation supported by the plot of logarithmic **1_3_spread** vs. logarithmic hardness in Figure 13, which demonstrates no obvious correlation between the two attributes.

V. CONCLUSIONS

Although an Intermediate class of pulses was indicated by this analysis, it is likely the result of sample incompleteness, and no definitive conclusions can be drawn about its properties. However, Short and Long pulse classes were much more definitively identified on the basis of pulse duration and spectral hardness. Short and Long pulses have different characteristics, and if they have otherwise similar characteristics, then the process that produces them occurs on a different timescale, which could represent a physical difference in how their progenitors deposit energy into the system.

The properties of Short and Long pulses found here are consistent with models for their formation, and with the properties of Short and Long bursts. In this way, the Short and Long burst classes were recovered through the use of pulse properties. As it has been shown that fewer pulses are needed to characterize GRBs than previously thought — in many cases, a single non-monotonic pulse effectively accounts for the observed features of the burst — pulse classes can be mapped almost

directly to burst classes, which is compelling evidence that pulses are the basic unit of GRB prompt emission.

Future work will be directed towards improving sample completeness with regards to multi-pulsed bursts, as well as increasing the overall number of pulses in the sample in order to more rigorously study differences within classes of Short and Long pulses.

Acknowledgements

Many thanks to Dr. Jon Hakkila, who motivated this projected and provided abundant guidance and advice, as well as Eric Hofesmann, who contributed to the development of the pulse sample. I would also like to thank Rebecca Brnich, Bailey Williamson, and Drew Ayers for their support and perspectives.

Notes and References

**Corresponding author email: hakkilaj@cofc.edu*

1. Klebesadel, R. W., Strong, I. B., \& Olsen, R. A. 1973, ApJ, 182, L85
2. Piran, T., 2004, arXiv:astro-ph/0405503
3. Norris, J. P., Nemiroff, R. J., Bonnell, J. T., et al. 1996, ApJ, 459, 393
4. Hakkila, J. and Preece, R.D., 2011, arXiv:1103.5434
5. Hakkila, J. and Preece, R.D., 2014, arXiv:1401.4047
6. Hakkila, J., Horvath, I., Hofesmann, E., & Lesage, S. 2018, ApJ, 855, 101
7. Mukherjee, S., et al. 1998, arXiv:astro-ph/9802085
8. Horvath, I. 1998, ApJ, 508, 757
9. Jordan, A., Hakkila, J., \& Nettles, C.J. 2010, AAS, 42, 229
10. Hakkila, J., et. al. 2003, arXiv:astro-ph/0209073
11. Scargle, J., Norris, J., \& Bonnell, J. 1997, arXiv:astro-ph/9712016
12. Scargle, J. 1998, arXiv:astro-ph/9711233
13. Hofesmann, E. 2016, "Characterizing the Complexity of GRB Light Curves," hofesmannen.stu.cofc.edu/files/GRB\%20Senior\%20Research.pdf
14. Norris, J.P., Bonnell, J. T., Kazanas, D. et al. 2005, arXiv:astro-ph/0503383